



TESIS

VALIDASI BIODATA MAHASISWA MELALUI PENDEKATAN NLP BERBASIS INDOBERT

FAJAR INDRA NUR ALAM
NPM 24066020006

DOSEN PEMBIMBING
Dr. Ir. I Gede Susrama Mas Diyasa, ST. MT. IPU.
Dr. Eng. Agussalim, S.Pd., MT.

KEMENTERIAN PENDIDIKAN TINGGI, SAINS, DAN TEKNOLOGI
UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR
FAKULTAS ILMU KOMPUTER
PROGRAM STUDI MAGISTER TEKNOLOGI INFORMASI
SURABAYA
2026

PENGAJUAN TESIS

VALIDASI BIODATA MAHASISWA MELALUI PENDEKATAN NLP BERBASIS INDOBERT

Tesis

Sebagai Salah Satu Syarat untuk Mencapai Gelar Magister Program Studi
Teknologi Informasi

Disusun dan diajukan oleh

FAJAR INDRA NUR ALAM
NPM. 24066020006

Kepada

FAKULTAS ILMU KOMPUTER
UNIVERSITAS PEMBANGUNAN “VETERAN” JAWA TIMUR
SURABAYA
2026

**VALIDASI BIODATA MAHASISWA MELALUI PENDEKATAN NLP
BERBASIS INDOBERT**

Disusun Oleh

FAJAR INDRA NUR ALAM

24066020006

Telah disetujui untuk mengikuti Ujian Tesis

Menyetujui,

Pembimbing I



Dr. Ir. I Gede Susrama Mas Diyasa,

ST. MT., IPU.

NIP. 198012052005011002

Pembimbing II



Dr. Eng. Agussalim, S.Pd., MT.

NIP. 198508112019031005

**Mengetahui
KOORDINATOR PRODI**



Dr. Eng. Agussalim, S.Pd., MT.

NIP. 198508112019031005

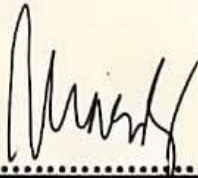
**VALIDASI BIODATA MAHASISWA MELALUI PENDEKATAN NLP
BERBASIS INDOBERT**

Oleh:
FAJAR INDRA NUR ALAM
NPM. 24066020006

Telah dipertahankan dihadapan dan diterima oleh Tim Penguji Ujian Tesis
Prodi Magister Teknologi Informasi Fakultas Ilmu Komputer Universitas
Pembangunan Nasional Veteran Jawa Timur
Pada tanggal 30 Juli 2026

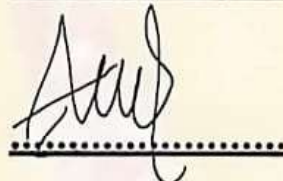
Menyetujui

**Dr. Ir. I Gede Susrama Mas Diyasa, ST. MT.,
IPU,**
NIP. 197006192021211009



(Pembimbing I)

Dr. Eng. Agussalim, S.Pd., MT.
NIP. 198508112019031005



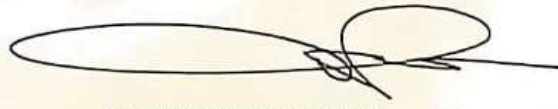
(Pembimbing II)

Dr. Faisal Muttaqin, S.Kom., MT
NIP. 198512312021211009



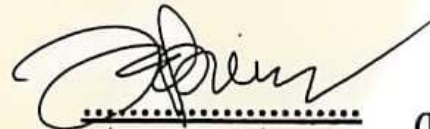
(Penguji I)

Prof. Dr. Basuki Rahmat, S.Si. MT
NPT. 196907232021211002



(Penguji II)

Dr. Rr. Ani Dijah Rahajoe, ST, M.Cs.
NPT. 197305122005012003



(Penguji III)

Mengetahui,
Dekan Fakultas Ilmu Komputer



Prof. Dr. Ir. Novirina Hendrasarie, MT
NIP. 19681126 199403 2 001

SURAT PERNYATAAN BEBAS PLAGIASI

Saya yang bertanda tangan dibawah ini :

Nama : Fajar Indra Nur Alam
NPM : 24066020006
Program : Magister (S2)
Program Studi : Magister Teknologi Informasi
Fakultas : Ilmu Komputer

Menyatakan bahwa dalam dokumen ilmiah Tesis ini tidak terdapat bagian dari karya ilmiah lain yang telah diajukan untuk memperoleh gelar akademik di suatu lembaga Pendidikan Tinggi, dan juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang/lembaga lain, kecuali yang secara tertulis disitasi dalam dokumen ini dan disebutkan secara lengkap dalam daftar pustaka.

Dan saya menyatakan bahwa dokumen ilmiah ini bebas dari unsur-unsur plagiasi. Apabila dikemudian hari ditemukan indikasi plagiat pada Tesis ini, saya bersedia menerima sanksi sesuai dengan peraturan perundang-undangan yang berlaku.

Demikian surat pernyataan ini saya buat dengan sesungguhnya tanpa ada paksaan dari siapapun juga dan untuk dipergunakan sebagaimana mestinya.

Surabaya, 30 Juli 2026
Yang membuat pernyataan



Fajar Indra Nur Alam
NPM. 24066020006

ABSTRAK

FAJAR INDRA NUR ALAM. *Validasi Biodata Mahasiswa Melalui Pendekatan NLP Berbasis IndoBERT* (dibimbing oleh **Dr. Ir. I Gede Susrama Mas Diyasa, ST. MT. IPU.** dan **Dr. Eng. Agussalim, S.Pd, MT.**).

Proses pendaftaran wisuda sarjana di perguruan tinggi kerap terhambat oleh ketidaksesuaian biodata mahasiswa antara pangkalan data akademik (SIAMIK) dan dokumen fisik ijazah Sekolah Menengah Atas (SMA). Pengecekan secara manual rentan terhadap human error dan memakan waktu yang lama. Penelitian ini bertujuan mengembangkan sistem cerdas untuk memvalidasi biodata pendaftaran wisuda secara otomatis guna meningkatkan efisiensi serta akurasi administrasi akademik. Metode yang diusulkan menggabungkan tiga pendekatan berkesinambungan: ekstraksi teks dari pindaian ijazah menggunakan antarmuka Vision-Language Model (Gemini API), pengenalan entitas bernama atau Named Entity Recognition (NER) memanfaatkan model bahasa IndoBERT yang di-fine-tuning pada 1.500 korpus data berformat CoNLL, dan pencocokan nilai string melalui algoritma Fuzzy String Matching (Token Set Ratio). Evaluasi kinerja sistem diukur melalui metrik Character Error Rate (CER) untuk akuisisi teks, evaluasi klasifikasi (Akurasi, Presisi, Recall, dan F1-Score) untuk NER, serta akurasi keabsahan pada validasi akhir. Hasil pengujian menunjukkan bahwa tahapan ekstraksi visual mencapai rata-rata CER yang sangat baik sebesar 4,2%. Hasil akurasi mencapai 99%. Pada tahapan pemahaman semantik, arsitektur IndoBERT membuktikan kinerja yang luar biasa dalam mengekstraksi entitas target (Nama Mahasiswa, Tempat Lahir, dan Tanggal Lahir) dengan nilai F1-Score pengujian mencapai 90%. Selanjutnya, modul pencocokan toleransi (Fuzzy Matching) terbukti mampu mengatasi variasi penulisan teks untuk memvalidasi kebenaran data ijazah terhadap referensi SIAMIK secara presisi berdasarkan ambang batas kemiripan 90%. Kesimpulannya, integrasi pendekatan pemrosesan bahasa alami berbasis IndoBERT secara signifikan berhasil meminimalkan intervensi manual, mengotomatisasi pengesahan dokumen, serta mengoptimalkan layanan pendaftaran wisuda sarjana.

Kata Kunci: Validasi Biodata, IndoBERT, Pendaftaran Wisuda, NER, Fuzzy Matching

ABSTRACT

FAJAR INDRA NUR ALAM. Student Biodata Validation Through an IndoBERT-Based NLP Approach (supervised by **Dr. Ir. I Gede Susrama Mas Diyasa, ST. MT. IPU.** and **Dr. Eng. Agussalim, S.Pd., MT.**).

The bachelor's degree graduation registration process in higher education is often hindered by discrepancies in student biodata between the academic database (SIAMIK) and the physical high school diploma document. Manual verification is susceptible to human error and is time-consuming. This research aims to develop an intelligent system to automatically validate graduation registration biodata to improve the efficiency and accuracy of academic administration. The proposed method combines three continuous approaches: text extraction from scanned diplomas using a Vision-Language Model interface (Gemini API), Named Entity Recognition (NER) utilizing the IndoBERT language model fine-tuned on 1,500 CoNLL-formatted data corpora, and string value matching through a Fuzzy String Matching algorithm (Token Set Ratio). System performance evaluation is measured through the Character Error Rate (CER) metric for text acquisition, classification evaluation (Accuracy, Precision, Recall, and F1-Score) for NER, and the validity accuracy at final validation. The test results show that the visual extraction stage achieves an excellent average CER of 4.2%. The accuracy result reached 99%. In the semantic understanding stage, the IndoBERT architecture demonstrates outstanding performance in extracting target entities (Student Name, Place of Birth, and Date of Birth) with a testing F1-Score reaching 90%. Furthermore, the tolerance matching module (Fuzzy Matching) is proven capable of handling text writing variations to precisely validate the correctness of diploma data against the SIAMIK reference based on a 90% similarity threshold. In conclusion, the integration of the IndoBERT-based natural language processing approach significantly succeeds in minimizing manual intervention, automating document authorization, and optimizing bachelor's degree graduation registration services.

Keywords: Biodata Validation, IndoBERT, Graduation Registration, NER, Fuzzy Matching

KATA PENGANTAR

Puji syukur kehadiran Allah SWT atas segala rahmat, hidayah dan karunia-Nya kepada penulis sehingga tesis dengan judul **“Validasi Biodata Mahasiswa Melalui Pendekatan Nlp Berbasis Indobert”** dapat terselesaikan dengan baik.

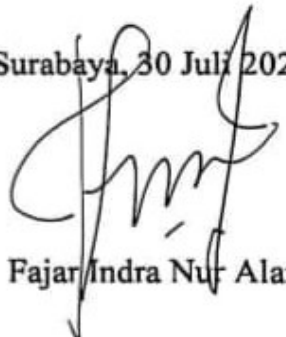
Peneliti menyadari tidak akan dapat menyelesaikan tesis ini dengan baik tanpa rida, bimbingan, saran, motivasi, dan bantuan dari berbagai pihak. Pada kesempatan ini, penulis mengucapkan terimakasih yang sebesar-besarnya kepada:

1. Ayah Tukiman, S.Sos, M.Si, dan Ibu Ngadiyah, S.Pd , selaku kedua orang tua yang selalu memberikan doa, dukungan moril maupun materiil, serta dukungan yang tiada henti.
2. Bapak Prof. Dr. Ir. Akhmad Fauzi, M.MT. selaku Rektor Universitas Pembangunan Nasional “Veteran” Jawa Timur.
3. Ibu Dr. Ir. Novirina Hendrasarie, MT. selaku Dekan Fakultas Ilmu Komputer Universitas Pembangunan Nasional “Veteran” Jawa Timur.
4. Ibu Dr. Eng. Agussalim, S.Pd., MT. Selaku Koordinator Program Studi Magister Teknologi Informasi Fakultas Ilmu Komputer Universitas Pembangunan Nasional “Veteran” Jawa Timur.
5. Bapak Dr. Ir. I Gede Susrama Mas Diyasa, ST. MT. IPU. selaku Dosen Pembimbing pertama, yang telah sabar memberikan bimbingan, motivasi, dan waktunya dalam proses bimbingan penyusunan tesis.
6. Bapak Dr. Eng. Agussalim, S.Pd., MT. selaku Dosen Pembimbing kedua, yang telah memberikan arahan dan saran dalam proses penyusunan tesis.
7. Bapak Dr. Faisal Muttaqin, S.Kom., M.Kom. selaku Ketua Penguji dan Dosen Wali, yang telah memberikan bimbingan, nasehat, kritik dan saran yang sangat berharga selama masa perkuliahan dan proses penyusunan tesis ini.
8. Bapak Prof. Dr. Basuki Rahmat, S.Si. MT. selaku Anggota Penguji, yang telah memberikan kritik dan masukan yang sangat bermanfaat dalam proses penyusunan tesis ini.

9. Ibu Dr. Rr. Ani Dijah Rahajoe, ST, M.Cs. selaku Anggota Penguji, yang telah memberikan kritik dan masukan yang sangat bermanfaat dalam proses penyusunan tesis ini.
10. Ibu Nurul Komariyah, S.Pd, M.M. dan Sunarto, S.Kom, selaku atasan tempat saya bekerja yang selalu memberikan motivasi menyelesaikan studi.
11. Hafizh dan Hanif selaku anak saya yang selalu memberikan saya doa, dukungan dan motivasi yang tiada henti.
12. Seluruh staf Tata Usaha dan civitas akademika Universitas Pembangunan Nasional “Veteran” Jawa Timur atas bantuan administrasi dan pelayanan selama penulis menempuh pendidikan.
13. Seluruh keluarga besar dan kerabat yang senantiasa memberikan doa dan dukungan.
14. Teman-teman seperjuangan yang selalu memberikan semangat, motivasi, dan kebersamaan selama proses penyusunan tesis.
15. Semua pihak yang tidak dapat disebutkan satu per satu yang telah membantu secara langsung maupun tidak langsung dalam penyusunan tesis ini.

Penulis menyadari bahwa di dalam penyusunan tesis ini banyak terdapat kekurangan. Untuk itu kritik dan saran yang membangun dari semua pihak sangat diharapkan demi kesempurnaan penulisan tesis ini. Akhirnya, dengan segala keterbatasan yang penulis miliki semoga laporan ini dapat bermanfaat bagi semua pihak umumnya dan penulis pada khususnya.

Surabaya, 30 Juli 2026



Fajar Indra Nur Alam

DAFTAR ISI

PENGAJUAN TESIS.....	iii
PERNYATAAN KEASLIAN TESIS DAN PELIMPAHAN HAK CIPTA	vi
ABSTRAK	viii
ABSTRACT	x
KATA PENGANTAR.....	xii
DAFTAR ISI.....	xiv
DAFTAR TABEL	xix
DAFTAR GAMBAR	xxi
BAB 1 1	
PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah.....	4
1.3 Tujuan Penelitian	4
1.4 Manfaat Penelitian	5
1.5 Batasan Masalah	6
BAB 2 7	
TINJAUAN PUSTAKA	7
2.1 Penelitian Terdahulu	7
2.2 Ijazah.....	12
2.3 Ekstraksi Teks Multimodal Berbasis <i>Large Language Model</i> (LLM)	12
2.3.1 Mekanisme Pemahaman Dokumen Visual secara Holistik.....	13
2.3.2 Rekayasa Instruksi Sebagai Pengendali Output.....	14
2.3.3 Rekayasa Instruksi (Prompt Engineering) dan Zero-Shot Inference)	14

2.3.4	Transisi Ekstraksi Menuju Downstream Task NLP	15
2.4	Praproses	16
2.4.1.	Praproses Akuisisi Berbasis Rekayasa Instruksi	16
2.4.2.	Praproses Teks Lanjutan.....	17
2.4.3	Penanganan Noise Visual dan Tekstual pada Ijazah.....	17
2.4.4	Interpolasi Lanczos (<i>Lanczos Resampling</i>).....	18
2.4.5	Enkode Base64 untuk Transmisi Data (<i>Base64 Encoding</i>).....	19
2.5	<i>Machine Learning</i>	19
2.5.1	Natural Language Processing.....	19
2.5.2	Artificial Neural Network	20
2.5.3	Softmax	21
2.5.4	Transfer Learning	22
2.5.5	Transformer	23
2.5.6	Optimzer AdamW	24
2.5.7	Loss Function Categorical Cross Entropy.....	25
2.5.8	Confusion Matrix	26
2.6	Model Bahasa IndoBERT	29
2.6.1	Mekanisme Tokenisasi dan Embedding.....	30
2.6.2	Mekanisme Fine-Tuning Untuk Klasifikasi Token	32
2.7	Named Entity Recognition (NER) dan Skema Pelabelan.....	35
2.7.1	Skema Pelabelan BIO (Begin, Inside, Outside)	36
2.8	Kerangka Kerja Antarmuka Web (Streamlit).....	37

BAB 339

METODOLOGI PENELITIAN 39

3.1	Desain Penelitian	39
-----	-------------------------	----

3.2	Tahapan Penelitian	40
3.3	Persiapan dan Studi Literatur	43
3.3.1	Lokasi Penelitian	43
3.3.2	Waktu Penelitian	44
3.3.3	Perangkat Penelitian (Lingkungan Komputasi)	44
3.4	Akuisisi Data	45
3.4.1	Dokumen Ijazah SMA (Data Tak Terstruktur)	46
3.4.2	Pangkalan Data SIAMIK (<i>Data Ground Truth</i>)	47
3.4.3	Prosedur Seleksi Dataset (<i>Sampling</i>)	47
3.5	Pra-Pemrosesan Data	48
3.5.1	Rasterisasi Dokumen (<i>PDF to Image Extraction</i>)	49
3.5.2	Standarisasi Dimensi Geometris (<i>Lanczos Resizing</i>)	49
3.5.3	Standarisasi Ruang Warna dan Kompresi Transmisi	50
3.6	Ekstraksi Teks (Optical Character Recognition) Menggunakan Gemini API	50
3.6.1.	Konstruksi Muatan Data (<i>Payload Construction</i>)	52
3.6.2.	<i>Eksekusi Zero-Shot Prompting</i>	52
3.6.3.	Manajemen Transmisi dan Penyimpanan Keluaran	53
3.7	Pelabelan Data	53
3.7.1	Skema BIO Tagging	54
3.7.2	Prosedur Anotasi	55
3.8	Pembagian Data	57
3.9	Skenario Ekstraksi dan Validasi Biodata	59
3.9.1	Pembacaan Teks Dokumen (Ekstraksi Multimodal)	59
3.9.2	Pengenalan Entitas dengan IndoBERT (<i>Named Entity Recognition</i>) ...	60

3.9.3	Pencocokan dan Toleransi Kesalahan (<i>String Matching</i>).....	60
3.10	Perancangan dan Pelatihan Model IndoBERT.....	61
3.10.1	Arsitektur Model Pengenalan Entitas (Core)	62
3.10.2	Proses Pelatihan dan Optimasi (<i>Fine-Tuning</i>).....	63
3.10.3	Penyelarasan Tokenisasi (Token Alignment)	64
3.10.4	Analisis Konvergensi dan Stabilitas Model	65
3.11	Evaluasi Hasil	66
3.11.1	Evaluasi Character Error Rate (CER)	69
3.11.2	Evaluasi Model Klasifikasi dan Validasi (IndoBERT).....	71
3.11.3	Evaluasi Kemiripan Teks (<i>Text Similarity</i>).....	72
3.11.4	Fungsi Keanggotaan (<i>Membership Function</i>).....	74
3.11.5	Mekanisme Pengambilan Keputusan (<i>Static Thresholding</i>)	76
3.12	Pengembangan Layanan Sederhana.....	77
3.12.1	Perancangan Alur Kerja (System Pipeline).....	77
3.12.2	Perancangan Antarmuka Pengguna	78
3.12.3	Integrasi Backend dan Keamanan Data	78
3.13	Spesifikasi Perangkat Keras dan Perangkat Lunak.....	79
3.13.1	Perangkat Keras (<i>Hardware</i>).....	80
3.13.2	Perangkat Lunak (<i>Software</i>).....	80
3.14	Jadwal Penelitian	80

BAB 4

HASIL DAN PEMBAHASAN 83

4.1.	Gambaran Umum Pengujian.....	83
4.1.1.	Lingkungan Eksperimen dan Dataset.....	83
4.1.2.	Implementasi Kelas Ekstraksi (Main Engine).....	84

4.2.	Proses Optical Character Recognition	85
4.3.	Proses Pelabelan Data	87
4.4.	Proses Perancangan Dan Pelatihan Model IndoBERT	88
4.5.	Proses Fuzzy String Matching dan Similarity	90
4.6.	Hasil Evaluasi Ekstraksi Teks	92
4.6.1.	Evaluasi Character Error Rate (CER)	92
4.6.2.	Ketahanan Ekstraksi terhadap Kualitas Dokumen Visual.....	95
4.7.	Hasil Pengenalan Entitas (IndoBERT NER).....	95
4.7.1.	Analisis Konvergensi Pelatihan Model	96
4.7.2.	Laporan Evaluasi Metrik	97
4.8.	Evaluasi Pencocokan <i>Fuzzy String</i> dan Kecepatan Pemrosesan.....	100
4.9.	Analisis Perbandingan dan Pembahasan.....	101
4.9.1.	Analisis Kesalahan Ekstraksi (<i>Error Analysis</i>)	101
4.9.2.	Peran <i>Fuzzy String Matching</i> dalam Validasi Akademik	102
4.9.3.	Relevansi Sistem <i>Hybrid</i> dalam Operasional Institusi.....	102
BAB 5		103
KESIMPULAN DAN SARAN		103
5.1	Kesimpulan	103
5.2	Saran	104
DAFTAR PUSTAKA.....		107
LAMPIRAN.....		115

DAFTAR TABEL

Tabel 2. 1 Perbandingan Penelitian Terkait (SOTA).....	9
Tabel 3. 1 Spesifikasi Label BIO Tagging untuk Ekstraksi Biodata.....	56
Tabel 3. 2 Rasio Data Latih dan Data Uji	58
Tabel 3. 3 Ringkasan Metrik Evaluasi Sistem	74
Tabel 3. 4 Alur Sistem.....	78
Tabel 3. 5 Jadwal Penelitian.....	81
Tabel 4. 1 Spesifikasi Lingkungan Implementasi dan Pengujian Sistem	84
Tabel 4. 2 Hasil Evaluasi <i>Character Error Rate</i> (CER) Berdasarkan Entitas	93
Tabel 4. 3 Log Konvergensi Nilai Kerugian (Loss) selama Pelatihan IndoBERT.....	96
Tabel 4. 4 Laporan Evaluasi Metrik Klasifikasi IndoBERT (<i>Fine-Tuned</i>)	97
Tabel 4. 5 Performa <i>Fuzzy Matching</i> dan Eksekusi Waktu.....	100

DAFTAR GAMBAR

Gambar 2. 1 Arsitektur <i>Deep Feed-Forward Neural Network</i> (Khadivi dkk., 2016)	20
Gambar 2. 2 Multi-class Confusion Matrix	28
Gambar 2. 3 Embedding pada BERT	32
Gambar 2. 4 Paradigma <i>Pre-training</i> dan <i>Fine-Tuning</i> pada Arsitektur BERT (Devlin dkk., t.t.)	33
Gambar 2. 5 pre-training dan Fine-tuning pada BERT	34
 Gambar 3. 1 Tahapan Penelitian	 41
Gambar 3. 2 Tahapan Pengumpulan Data	45
Gambar 3. 3 Gambar Format data pdf ijazah SMA dan Biodata pada siamik csv	46
Gambar 3. 4 Flowchart Pra Pemrosesan Citra	48
Gambar 3. 5 Alur Kerja Gemini API 3.1 Flash Lite	51
Gambar 3. 6 Alur Ekstraksi Teks dengan Gemini API	52
Gambar 3. 7 Flowchart Pelabelan Data	53
Gambar 3. 8 Prosedur Anotasi BIO Tagging	55
Gambar 3. 9 Flowchart Pembagian Data	57
Gambar 3. 10 Flowchart perancangan dan pelatihan model IndoBERT	62
Gambar 3. 11 Block diagram Proses Fine Tuning IndoBERT untuk NER	66
Gambar 3. 12 Flowchart tahap evaluasi hasil	67
Gambar 3. 13 Flowchart Fuzzy String Matching	73
Gambar 3. 14 Tampilan Layanan Sederhana	79
 Gambar 4. 1 Arsitektur Pipa Pemrosesan Hibrida pada Kelas Main Engine	 85
Gambar 4. 2 Proses OCR menggunakan Gemini API 3.1 Flash Lite	86
Gambar 4. 3 OCR Result Gemini 3.1. Flash Lite	86
Gambar 4. 4 Proses Pelabelan Data dengan label studio	87
Gambar 4. 5 Pemetaan token dan label BIO Format CoNLL	88
Gambar 4. 6 Model Sebelum Fine Tuning dan sesudah fine tuning	89

Gambar 4. 7 Proses Fuzzy String Matching 100% mencocokkan hasil ekstraksi dengan data pada siamik.....	91
Gambar 4. 8 Proses Fuzzy String Matching belum 100% mencocokkan hasil ekstraksi dengan data pada siamik	92
Gambar 4. 9 Gambar Evaluasi Metrik Klasifikasi IndoBERT (<i>Fine-Tuned</i>)	98